130

Reihe Ökonomie

Economics Series

Testing for Relative Predictive Accuracy:
A Critical Viewpoint

Robert M. Kunst

130

Reihe Ökonomie

Economics Series

Testing for Relative Predictive Accuracy:
A Critical Viewpoint

Robert M. Kunst

May 2003

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Contact:

Robert M. Kunst
Department of Economics and Finance
Institute for Advanced Studies
Stumpergasse 56
1060 Vienna, Austria
☎: +43/1/599 91-255
fax: +43/1/599 91-163
email: kunst@ihs.ac.at

Founded in 1963 by two prominent Austrians living in exile – the sociologist Paul F. Lazarsfeld and the economist Oskar Morgenstern – with the financial support from the Ford Foundation, the Austrian Federal Ministry of Education and the City of Vienna, the Institute for Advanced Studies (IHS) is the first institution for postgraduate education and research in economics and the social sciences in Austria. The **Economics Series** presents research done at the Department of Economics and Finance and aims to share "work in progress" in a timely way before formal publication. As usual, authors bear full responsibility for the content of their contributions.

Das Institut für Höhere Studien (IHS) wurde im Jahr 1963 von zwei prominenten Exilösterreichern – dem Soziologen Paul F. Lazarsfeld und dem Ökonomen Oskar Morgenstern – mit Hilfe der Ford-Stiftung, des Österreichischen Bundesministeriums für Unterricht und der Stadt Wien gegründet und ist somit die erste nachuniversitäre Lehr- und Forschungsstätte für die Sozial- und Wirtschafts-wissenschaften in Österreich. Die **Reihe Ökonomie** bietet Einblick in die Forschungsarbeit der Abteilung für Ökonomie und Finanzwirtschaft und verfolgt das Ziel, abteilungsinterne Diskussionsbeiträge einer breiteren fachinternen Öffentlichkeit zugänglich zu machen. Die inhaltliche Verantwortung für die veröffentlichten Beiträge liegt bei den Autoren und Autorinnen.

Abstract

Tests for relative predictive accuracy have become a widespread addendum to forecast comparisons. Many empirical research reports conclude that the difference between the entertained forecasting models is 'insignificant'. This paper collects arguments that cast doubt on the usefulness of relative predictive accuracy tests. The main point is not that test power is too low but that their application is conceptually mistaken. The features are highlighted by means of some Monte Carlo experiments for simple time-series decision problems.

Keywords

Information criteria, forecasting, hypothesis testing

JEL Classifications

C12, C32, C53

Contents

1 Introduction

> *There has been too much formalism, tradition, and confusion that leads people to think that statistics and statistical science is mostly about testing uninteresting or trivial null hypotheses, whereas science is much more than this. We must move beyond the traditional testing-based thinking because it is so uninformative* [Burnham and Anderson, 2002, p.42]

A decade ago, comparative studies of the predictive performance of time-series models were usually presented on the basis of lists of descriptive statistics such as mean squared errors or their ratios across models. The contribution of Diebold and Mariano (DM) has revolutionized this practice. In the late 1990s, hardly any forecasting study was published in a major academic journal without using their test or one of its later refinements. A typical view is the one expressed by Fildes and Stekler (2002, p.439) in a recent survey paper "*Whatever benchmark is used in the evaluation of forecasts, the difference between the two sets of errors should be tested for statistical significance*" [original italics]. The main aim of this paper is to express some caution regarding such "should be" prescriptions.

In short, the DM test is based on the following ideas. A forecaster, who does not know the data-generation mechanism of given time-series data, entertains a set of models—using 'model' throughout in the sense of a parameterized collection of probability distributions—and compares their forecasting performance on a part of the sample, typically the most recent segment. A positive transform of the prediction errors serves as a moment or cost function. One of the entertained models is chosen as a baseline model. The forecaster considers the null hypothesis that a given model is unable to improve predictive performance relative to that model. Only if the DM statistic rejects this null hypothesis, can the competing and more sophisticated model definitely be recommended.

Even at a first glance, some informal arguments can be raised against this testing strategy. Everyone who has worked in professional forecasting will know that the cost of using a more sophisticated model is small. Just to the contrary, administrative directors of forecasting institutions may actually prefer a sophisticated model over a simple one, as such choice will improve the reputation of the institution. Therefore, the forecaster is not in the situation of classical hypothesis testing. There is no need to be conservative and there is no coercive assignment of null models or null hypotheses. Rather, the forecaster is in a decision situation. The best workhorse among a group of models has to be selected. The appropriate statistical framework is not hypothesis testing but rather model selection. Appropriate methods for model selection can be found in information theory and AIC–type criteria, or in Bayesian posterior-odds analysis. These methods are tuned to make a specific selection from a finite set, while hypothesis testing implies an interval of 'rejection failure', within which some models cannot be ranked. Such a 'demilitarized zone' does not appear to be a useful innovation but rather constitutes a practical inconvenience. Particularly

for DM testing, this interval appears to be rather wide in many applications, as will be demonstrated in this paper.

Moreover, while the DM test formally does not take refuge to the concept of a 'true model', its very null hypothesis reflects that concept. From a Bayesian viewpoint, classical hypothesis testing is justified only if the sharp or point null can be assigned a non-zero prior weight. However, the null hypothesis of the DM test is *a priori* improbable. Economic reality is usually seen as a complex dynamically evolving structure whose hidden time-constant laws are almost impossible to retrieve, in the spirit of the 'Haavelmo distribution'. Because none of the entertained forecasting models comes anywhere near this complexity, different models are just different approximations and imply different distributional properties of their prediction errors. Although certain moments may coincide across these distributions by pure chance, it is difficult to imagine assigning a non-zero prior weight to this event.

This paper first focuses on such theoretical aspects of significance tests for predictive accuracy. From a statistical viewpoint, it is quite difficult to justify their usage. However, it may still be convenient to apply the tests from an empirical perspective. The implicit strengthening of the prior weight on simple models could be favorable for model selection decisions in certain situations, even if it were untenable from a theoretical viewpoint. Monte Carlo simulations will serve to assess this argument. The simulations are based on simple time-series models and analyze the performance of model selection guided by DM tests in some nested and also non-nested situations. The truly empirically relevant case may be much more complex, as the typical forecaster uses a handful of simple ideas to model an economic reality far beyond the reach of parsimoniously parameterized structure. Nevertheless, the presented simulations point to some recurrent features. Firstly, the demilitarized zone generated by DM tests prevents any useful model choice even in situations where such a model choice has quite clear support from usual model selection statistics. Secondly, if the aim is selecting the true model or even pseudo-true model structure, comparing measures of predictive accuracy is a poor substitute for using information criteria, whether DM tests are used or not. Thirdly, however, the situation is less clear if the aim of selecting the optimum prediction model replaces the quest for a true structure. Prediction criteria may be more likely to hit upon the optimum prediction model than information criteria, and, moreover, the simplicity bias that is characteristic for DM testing may even improve upon this choice.

As a bottom line, a final answer to the question of whether to use DM tests or not requires clearly stating the aims of modeling. It is indeed unfortunate that the econometric and the forecasting literature alike have done little to separate the targets of searching 'true' models and of optimizing prediction and have repeatedly tended to blur the distinction between these aims. This is evident from the exaggerated concern about 'model misspecification' in a forecasting situation—where misspecified models may yield excellent forecasts—and from statements like "*it is our fervent belief that success in* [the evaluation and improvement of forecasting performance] *should also lead to an improved understanding of the economic process*" [Fildes and Stekler, 2002, p. 462,

original italics]. Taking into account that good forecasting models may be poor or incorrect descriptions of the data-generating mechanism—even in situations where such a data-generating mechanism *exists*—and *vice versa* may suggest to regard such paradigms with the utmost caution.

The plan of the paper is as follows. Section 2 reviews the DM test statistic. Section 3 explores some basic properties of this statistic in small samples—throughout the paper, a sample size of $n = 100$ is used—against the backdrop of time-series model selection among AR(1), MA(1), and ARMA(1,1) structures when the true generating model is ARMA(1,1) with positive coefficients. Section 4 discusses some theoretical arguments against the usage of significance checks on model selection decisions. Section 5 explores the performance of predictive accuracy evaluations as model selection criteria, both from the viewpoint of finding true structures and of finding the best forecasting model In Section 5.1, the standard design of ARMA(1,1) is maintained, while Section 5.2 repeats the simulation experiments for designs with random coefficients. Section 6 concludes.

Here, an important note is in order. This paper does not aim at criticizing the literature for any lack of correctness, particularly not the work of DM and of Linhart. Neither does it focus on problems of the lack of test power in the considered procedures, particularly as various modifications of the original DM test have been suggested recently. The critique rather focuses on the methodological concept of the tests at a fundamental level. For a recent critique of DM test power, see Ashley (in press).

2 The Diebold-Mariano test

It is useful to review the original procedure as it was suggested by Diebold and Mariano (1995, DM). DM motivated their contribution in the following paragraph:

> *Given the obvious desirability of a formal statistical procedure for forecast-accuracy comparisons, one is struck by the casual manner in which such comparisons are typically carried out. The literature contains literally thousands of forecast-accuracy comparisons; almost without exception, point estimates of forecast accuracy are examined, with no attempt to assess their sampling uncertainty. On reflection, the reason for the casual approach is clear: Correlation of forecast errors across space and time, as well as several additional complications, makes formal comparison of forecast accuracy difficult.*

While DM do not really specify the 'additional complications', some of these are outlined in the remainder of their paper, such as non-normality and small samples. In this paper, we contend that the 'casual manner' may be preferable to the 'obviously desired' testing approach, which argument will be supported by some simulations.

Contrary to what is usually cited as 'the DM test', DM suggest various testing procedures with similar aims and subject them to some Monte Carlo comparisons. In its narrow sense, the DM statistic appears to be the statistic S_1, which is introduced as an 'asymptotic test' for the null hypothesis that $\mathrm{E}(d_t) = 0$, when $d_t = g(e_{jt}) - g(e_{it})$, with e_{it} and e_{jt} forecast errors from using two different forecasting procedures that could be indexed i and j. The function $g(.)$ is a loss function. Although DM do not rigorously specify the properties of $g(.)$, it is reasonable to assume that $g(.)$ be positive and that $g(x) = 0$ for $x = 0$ only. Typical loss functions are $g(x) = |x|$ for mean absolute errors (MAE) and $g(x) = x^2$ for mean squared errors (MSE). It may be useful to also consider certain cases of asymmetric or bounded loss functions, though some monotonicity in the form of $g(x) \geq g(y)$ for $x > y > 0$ or $x < y < 0$ may be reasonable.

With these definitions, DM consider the statistic $\bar{d}$ defined as the time average over a sample of d_t, $t = 1, \ldots, n$, $\bar{d} = n^{-1} \sum_{t=1}^{n} d_t$. It is easily shown that the standardized statistic

$$S_1 = \frac{\bar{d}}{\sqrt{n^{-1} 2\pi \hat{f}_d(0)}} \tag{1}$$

converges to a standard normal distribution for $n \to \infty$. The element in the denominator $\hat{f}_d(0)$ is a consistent estimator of the spectral density of d_t at the frequency 0, such as

$$\hat{f}_d(0) = (2\pi)^{-1} \sum_{k=-n+1}^{n-1} w(k/S(n)) \hat{\gamma}_d(k), \tag{2}$$

with the lag window function $w(.)$, the truncation lag $S(n)$, and the estimated autocorrelation function $\hat{\gamma}_d(.)$. Typical choices for $w(.)$ would be the Bartlett window or the rectangular window. While DM favor the rectangular window, the Bartlett window appears to be the better choice for small n, as it does not require to choose a too small value for $S(n)$.

Apart from the S_1 statistic, DM also consider some non-parametric tests, in particular for the case where 'only a few forecast-error observations are available', which may be the empirically relevant case. Apart from rank tests and sign tests, they review a test following Meese and Rogoff (1988), which checks the correlation of $e_{it} - e_{jt}$ and $e_{it} + e_{jt}$ and whose idea may also be generalized to the loss-function framework of the S_1 construction. In the remainder of this paper, the standard normal test based on the statistic S_1 will be regarded as 'the DM test', with $w(.)$ specified as the Bartlett window.

3 Some basic properties of DM testing in action

3.1 The basic simulations

Some simulations may serve to highlight the main features at stake. The predictive performance of some rival models is evaluated, some of which are 'mis-

specified'. For the baseline simulations, 1000 replications of time series of length 210 are generated. The data-generating process is a simple ARMA(1,1) process

$$x_t = \phi x_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} \tag{3}$$

with ε_t independently drawn from a $N(0, 100)$ distribution. The coefficients ϕ and θ are varied over the positive unit interval $[0, 1]$ in steps of 0.05. The first 100 observations of each trajectory are discarded. Simple time-series models, such as AR(1), MA(1), and the 'true' model ARMA(1,1) are fitted to the observations 101 to 200 in order to provide parameter estimates $\hat{\theta}$ and $\hat{\phi}$. These parameter estimates and the observations 101 to 200 are then used to generate forecasts for the observation 201. The prediction error is stored. This out-of-sample forecasting is repeated for the estimation range 102 to 201 to predict the observation 202, and another prediction error is stored. This rolling out-of-sample forecasting is repeated until the end of the generated sample is reached. Rolling out-of-sample forecasting is a widespread technique in the econometric literature (see, e.g., TASHMAN, 2000). Thus, each model generates 10,000 prediction errors for each fixed parameter (ϕ, θ), from which moment statistics are calculated, such as the mean absolute error (MAE) or the mean squared error (MSE). The reported results focus on the MAE. Note that all parameter estimates rely on exactly 100 observations.

This experiment is not new and presumably has been reported by other authors. However, it will serve as an instructive background. It is an experiment in the classical statistics framework, as all simulations are conducted conditional on fixed and true parameter values. Figure 1 shows the difference MAE(MA)-MAE(AR) in the form of a contour plot. Unless $\phi = 0$ or $\theta = 0$, both forecasting models are 'misspecified'. It is seen that the MA forecast is preferable because of its smaller MAE for the MA processes, but also for AR processes with $\phi < 0.15$ and for some mixed processes with $\theta > \phi$. In the area to the right, AR models yield the more precise prediction. The figure allows crude recommendations to empirical forecasters who, for technical reasons, do not want to use mixed models for forecasting. For example, if the data support an ARMA(0.8,0.8) model, autoregressive forecasts are preferable to moving-average forecasts.

A different picture emerges when the MA(1) forecast model is compared to an ARMA(1,1) model. The ARMA(1,1) is the true model for all trajectories and should dominate the MA(1) model for all parameter values asymptotically, excepting $\phi = 0$. Figure 2 shows that for $\phi < 0.2$, the MA(1) model yields better forecasts if these are assessed by the MAE. For larger ϕ, the precision of MA forecasts deteriorates quickly, which is indicated by the steepness of the slope. Contour curves are shown in the range $[-1, 1]$. Even in the area to the left of the MAE(MA)=MAE(ARMA) curve, the difference in accuracy never exceeds 0.1, hence the loss to the forecaster from using the ARMA model in all cases appears to be small. By contrast, the MA model yields relatively poor predictions for $(\phi, \theta) = (0.6, 0.9)$, where the MA forecast still dominates the AR forecast, as can be seen from Figure 1.

The counterpart for the AR(1) versus the ARMA(1,1) model is shown in Figure 3. Its symmetry to Figure 2 is surprisingly exact, although the increase

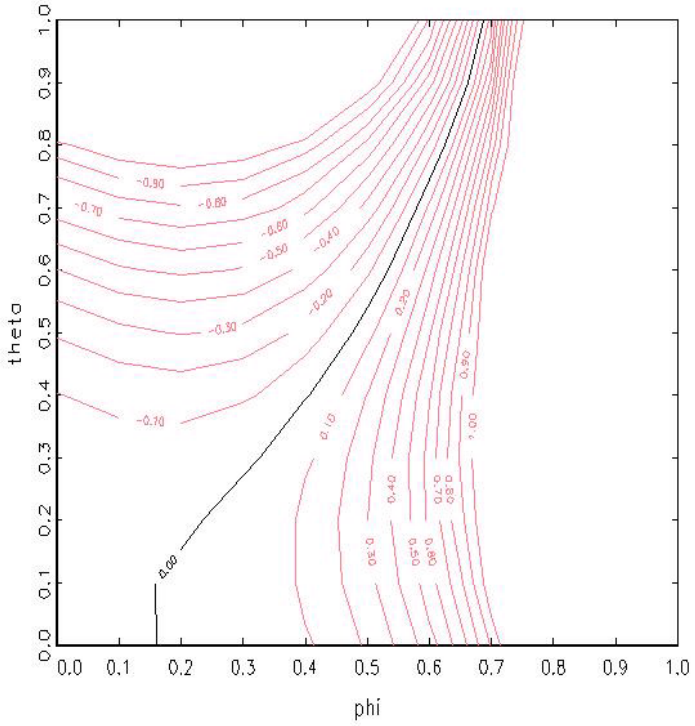

Figure 1: MAE for the MA(1) forecast model minus MAE for the AR(1) forecast model when the true model is ARMA(ϕ, θ). Sample size is 100.

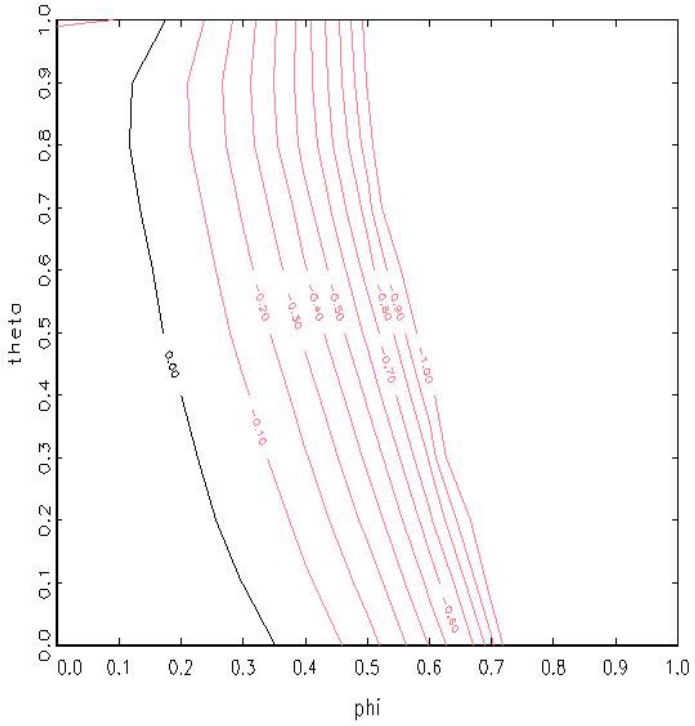

Figure 2: MAE for the MA(1) forecast model minus MAE for the ARMA(1,1) forecast model when the true model is ARMA(ϕ, θ). Sample size is 100.

of the MAE of the AR(1) forecast as θ rises is markedly slower than for the MA(1) forecast as ϕ rises. The values for $(\phi, \theta) = (1, 1)$ should be ignored, as they may only reflect numerical instabilities of the maximum-likelihood estimate of the GAUSS routine. The three figures can be superimposed and demonstrate that the ARMA(1,1) model is to be preferred roughly whenever $\phi > 0.2$ *and* $\theta > 0.2$. In the band close to the axes, the simpler models AR(1) and MA(1) dominate, with an asymmetric preference toward MA(1) in the southwest corner. The collected evidence from the simulations allows useful empirical guidelines for the forecasting practitioner.

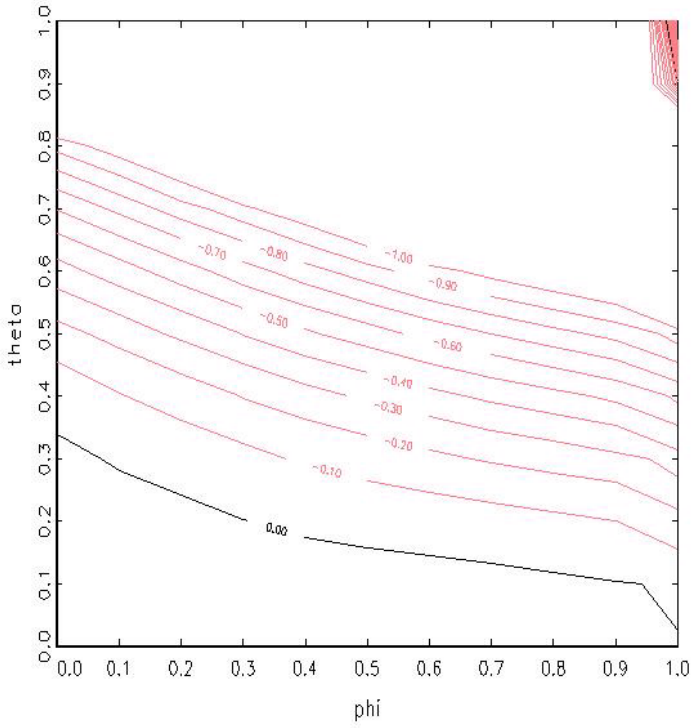

Figure 3: MAE for the ARMA(1,1) forecast model minus MAE for the AR(1) forecast model when the true model is ARMA(ϕ, θ). Sample size is 100.

3.2 Some simulations with Diebold-Mariano testing

For these simulations, the data-generating process is identical to that used in the previous subsection. After estimation and forecasting, i.e., after conducting the ten one-step predictions, a DM test statistic is calculated and is compared to the two-sided 0.05 asymptotic significance point, i.e., the 0.975 fractile of the standard normal distribution. Like all significance levels in frequentist testing, this

choice is debatable. The intention was to impose a relatively strict significance level in order to underscore the effects without biasing the results unduly, as for example by an 0.01 level. 1,000 replications allow to calculate the frequency of rejections and non-rejections of the null hypothesis that $\mathrm{E}\,|\,|e_{1,t}| - |e_{2,t}|\,| = 0$. Figure 4 shows that the frequency of the finding that autoregressive forecasts are 'significantly' better than moving-average forecasts exceeds 0.5 only if $\phi > 0.9$, whereas moving-average forecasts only achieve a rejection frequency of 0.4 in the extreme north-west corner. In other words, for all ARMA processes except for the nearly integrated ones, the researcher will find on average that there is no significant difference between the two forecasting models. Assuming a uniform prior distributions over the unit square in the Bayesian style, one may even conclude that the researcher will be unable to make a decision in 90% of the possible cases.

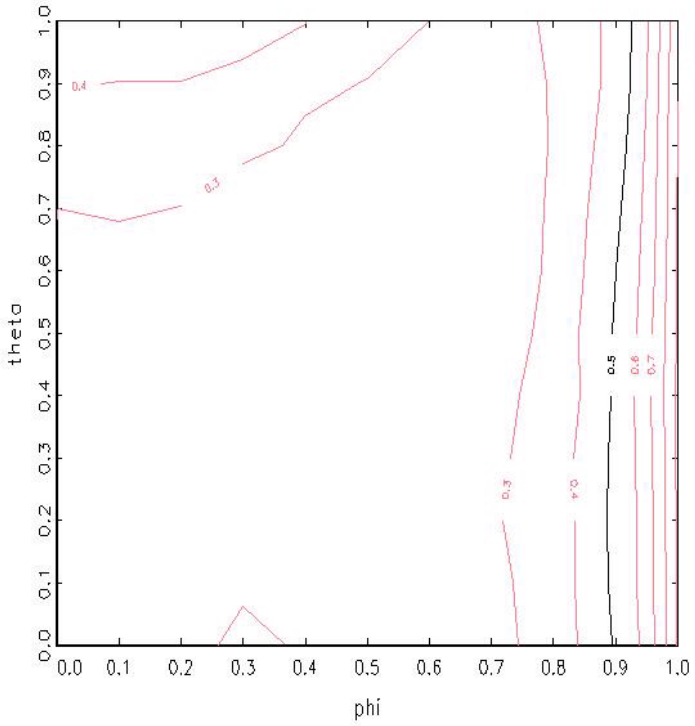

Figure 4: Frequency of a rejection of the DM test for MA(1) versus AR(1) predictions, if the true model is ARMA(1,1). 1,000 replications of series of length 100, with 10 single-step predictions for each trajectory.

Figure 5 shows a comparable experiment for the two rival models AR(1) and ARMA(1,1). Generally, the difference between the two forecasting models is 'insignificant', with rejection rates of the DM statistic ranging from 0.2 to 0.5.

The null hypothesis of equal forecasting precision is also rejected at least in 20% of the simulated cases even for autoregressive models with $\theta = 0$, where the AR forecasts are definitely more efficient, as no additional parameter is estimated. On the other hand, the ARMA forecast is 'significantly better' only for $\phi > 0.3$ and $\theta > 0.9$. In other words, the DM test is unable to recommend the ARMA model for most parameter constellations, even when the generating model has a substantial moving-average component.

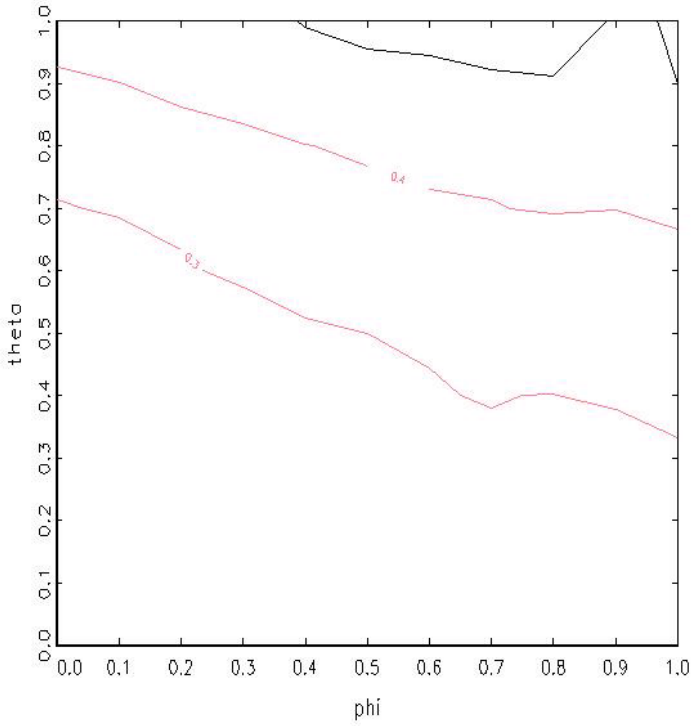

Figure 5: Frequency of a rejection of the DM test for AR(1) versus ARMA(1) predictions, if the true model is ARMA(1,1). 1,000 replications of series of length 100, with 10 single-step predictions for each trajectory.

Finally, Figure 6 shows the experiment for the rival models MA(1) and ARMA(1,1). This is a nested situation, the nesting model being the true one for all cases excepting $\phi = 0$. Only for $\phi > 0.9$ does the DM test imply a decision in favor of the ARMA(1,1) model with a probability of more than 0.5. For $\phi < 0.9$, the incorrect parsimonious model and the true structure yield forecasts of a comparable quality, at least if one believes in the results of the DM test. Consulting Figure 2, it is seen that the true model gives better forecasts indeed even for $\phi > 0.25$.

The experiments of this section demonstrate that the DM test can be a quite

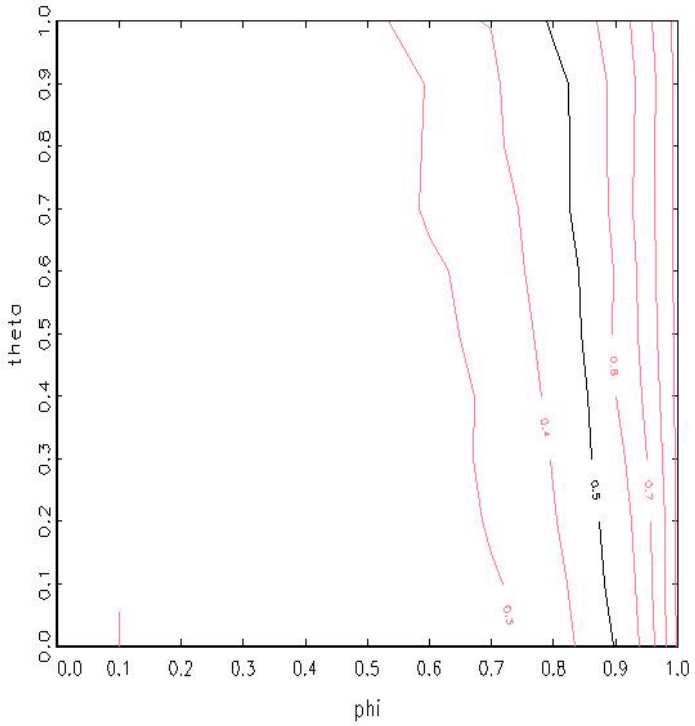

Figure 6: Frequency of a rejection of the DM test for MA(1) versus ARMA(1) predictions, if the true model is ARMA(1,1). 1,000 replications of series of length 100, with 10 single-step predictions for each trajectory.

useless instrument, if a decision concerning the forecasting model is searched for. This conclusion is valid for nested situations as well as for the more realistic non-nested cases. Such experiments could be extended in several directions. Firstly, the training interval can be extended and the estimation interval shortened. A certain ratio of these lengths will optimize the accuracy of decision even if based on the DM test only. Secondly, noting that the informative Figures 1–3 are obtained from decisions based on 1,000 replications while the non-informative Figures 4–6 reflect decisions based on single trajectories, it may be interesting to assess the relative merits of extending the sample length as compared to sampling additional trajectories. It may be conjectured that sampling additional trajectories is the most promising route. In practice, the relative merits of a certain forecasting model can only be assessed by considering a large number of comparable data sets. Such gathering of cases cannot be substituted by additional sophistication within one case. For example, the question whether model A or B are preferable for forecasting gross industrial output will remain undecided if DM tests are used. It may lead to an erroneous conclusion if descriptive horse races are evaluated. Repeating such descriptive horse races for a large number of comparable economies and averaging or summarizing the results will enable a more accurate decision, however, while this collection of parallel evidence does not appear to be encouraged by focusing on DM statistics for single cases.

It is worth while noting that the performance of the DM–based model selection procedure can be improved by changing the significance level from 0.05 to 0.1 or even 0.2. Because standard hypothesis testing gives no prescription on the way that significance levels depend on the sample size, such amendments remain arbitrary. The original contribution of DM focuses on the significance level of 0.1 exclusively, with sample sizes of prediction intervals ranging from 8 to 512, thus containing the value of 10 that is used in this paper.

4 The dangers of double checks

> *The practical utility of hypothesis testing procedures is of limited value in model identification.* [Akaike 1981, p. 722]

The contributions of Akaike (1974) and Schwarz (1978) started a revolution in time-series model selection. Rival methods, such as the visual inspection of sample moment functions recommended by Box and Jenkins (1976) and later Tiao and Tsay (1984) and nested sequences of hypothesis tests, lost the market of core time-series analysis to the information-criterion (IC) approach. It is interesting that empirical economics has remained largely in a pre-revolutionary stage, with a preference for classical testing within the Neyman-Pearson framework. The hidden controversy can be seen from comparing current standard textbooks on time series, such as Brockwell and Davis (2002), the attempted synthesis in an introductory textbook on econometrics by Ramanathan (2002), the slightly misplaced comments on IC in an

advanced econometrics textbook by GREENE (1990), and the radical omission of IC in an economic time-series book by HAMILTON (1994). On the battlefield of such an undeclared warfare, it may well be that an empirical paper gets rejected by a statistical journal due to the non-usage of IC, while adding them in a revision and submitting the manuscript to an economic journal will cause another rejection, as the information criteria 'although intuitively appealing, ... have no firm basis in theory' (GREENE, p.245). Although economists may not be aware of this background, the warfare on IC versus significance testing has a long history in the statistical discipline. It is a new version of the fight of Bayesian and frequentist paradigms.

In the frequentist paradigm, methods to tackle the finite-action problem directly have never been developed. Rather, model selection was embedded in the framework of testing a null hypothesis against an alternative. Because hypothesis testing is geared to binary and nested decisions, it can be applied to finite action only by means of sequences of test decisions and of artificial nesting. A single element in a frequentist testing chain mainly relies on a calculation of a measure of divergence between likelihoods, which are maximized for the two rival hypotheses, and the evaluation of the tail properties of the distribution of the divergence statistic, assuming the 'restricted' model to be correct. As in many frequentist tests, the significance level, i.e., the quantile of the distribution, remains arbitrary. As in all sequential tests, the arbitrariness of this significance level plays a crucial role and severely affects the final outcome of the test chain. The frequentist reluctance to eliciting prior distributions on parameter spaces prevents any systematic optimization of significance levels, beyond problematic aims such as 'keeping the overall significance level at 5%'.

Even in classical texts on regression analysis, purely descriptive statistics are reported side by side with test statistics for significance assessment. For example, the 'regression F' is viewed as a test statistic for checking the null hypothesis that all regression coefficients are zero, while the 'coefficient of determination' R^2 is viewed as a descriptive statistic, even though the former and the latter statistic are linked by a one-one transform. In its descriptive interpretation, the R^2 commonly serves to informally assess the descriptive quality of a model and to compare rival linear regression models. Note that comparative statistics, such as ratios of MSE or MAE across models, perform a similar function in forecasting experiments, or at least did so before DM tests were introduced.

The large-sample properties of model selection on the basis of an R^2 and even of its adjusted version due to WHERRY-THEIL are well known and generally discourage its usage for this purpose. By contrast, the optimization of IC variants was shown to lead to true structures (BIC) or otherwise preferable models (AIC, FPE). Like the R^2, IC statistics can be used for all model comparisons, including the situation of non-nested rival models. Contrary to a hypothesis testing decision, the IC decision is discrete and apparently non-stochastic, as preference corresponds to a simple operation of binary mathematical operators such as $<$ and $>$. Similarly, the implied MSE can be compared across forecasting models by a simple and exact comparison. This correspondence is not a mere coincidence. Roughly, optimization of the out-of-sample MSE approximates the

optimization of the AIC, as can be seen from the original derivation of the AIC criterion.

As a statistic calculated from data that are interpreted as realizations of random variables, the AIC has a statistical distribution that depends on the data-generating process. Therefore, the 'significance' of an AIC decision can be derived in principle. This approach was considered by LINHART (1988). This and similar contributions were generally ignored by the time-series literature, and with a good reason. By construction, the AIC decision removes the frequentist emphasis on a significance level and thus achieves an automatic and implicit adjustment of such significance level.

In a sense, the contribution by DIEBOLD AND MARIANO (1995) is the equivalent to the one by LINHART (1988). Where LINHART implicitly imposes a secondary significance test on a primary decision according to the AIC, DIEBOLD AND MARIANO impose a secondary significance test on a primary decision according to a summary measure of prediction errors. Unlike LINHART's idea, the suggestion by DIEBOLD AND MARIANO was welcomed immediately by the forecasting research community and it was widely applied. This remarkable divergence in reception by the academic community can tentatively be explained by two main arguments: firstly, the evaluation of prediction errors was not recognized as a decision procedure; secondly, the dominance of frequentist over Bayesian statistics is more pronounced in econometrics than in other fields of statistical applications. The analogy of predictive evaluation and of AIC evaluation will be demonstrated by some more simulation experiments in the next section.

A maybe too simplistic example of a possible 'double check' is the following. Suppose somebody wishes to test for the null hypothesis $H_0 : \mu = 0$ in a sample of Gaussian random observations with known unit variance against the alternative $H_A : \mu \neq 0$. On a significance level of 0.05, decision is usually based on a mean statistic $S_0(n) = n^{-1/2}\bar{x}$, where $\bar{x}$ denotes the sample mean and n is the sample size. For $|S_0(n)| > 1.96$ the null hypothesis is rejected, otherwise H_0 is retained. This decision is based on the property that $S_0(n)$ will converge in distribution to the $N(0, 1)$ law. However, for a given infinite random realization, $S_0(n)$ converges to a limit point that can be depicted as the integral over a Brownian motion trajectory. One may then be interested in whether this limit point $S_0(\infty) = \lim_{n\to\infty} n^{-1/2}\bar{x} = \lim_{n\to\infty} S_0(n)$ justifies a rejection. In this sense, the null hypothesis of interest will be $\tilde{H}_0 : |S_0(\infty)| < 1.96$ or, in words, that 'μ is insignificantly different from 0'. Because $S_0(n)$ shows some sampling variation around its limit $S_0(\infty)$ for finite n, a much higher value than 1.96 will be needed to reject $\tilde{H}_0$. The testing hierarchy can even be continued. Of course, the results of the secondary test on $\tilde{H}_0$ are altogether uninteresting and contradict basic statistical principles. However, that secondary test is a simple analogue to LINHART's test on the AIC and to the DM test on the significance of relative predictive accuracy.

To justify double-checking, it is argued in the literature that the researcher must safeguard against unnecessary complexity. A quote from LINHART (1988) is revealing: "Model 7 is less complex and one could be of the opinion that it

should be preferred to model 2 unless the hypothesis that model 2 is worse can be rejected at a small significance level" Indeed, this is exactly what information criteria are designed for, as they penalize the complex model 2 such that it has to beat the model 7 convincingly in order to be chosen. LINHART's test is unable to provide any new evidence in favor of neither model 2 nor model 7. It can only implicitly increase the penalty for complexity. In other words, the LINHART test counteracts the tendency of AIC to prefer complex models too often. It is not known whether the combined AIC–LINHART procedure achieves the full consistency of SCHWARZ' BIC for chains of tests in pure autoregressions. In small samples, it appears that AIC–LINHART is too conservative and chooses too parsimonious models.

A similar observation may hold with respect to the DM test and model selection by forecasting accuracy. Like AIC, out-of-sample forecasting accuracy criteria contain a penalty for spurious complexity that is typically so high that the true model class is ruled out due to sampling variation in parameter estimation, as was demonstrated in Figures 1–6. The only effect that can be achieved by a double-check test is to bias selection even more and probably unduly in the direction of the simple structures. This undue bias implies just what the forecaster wants to avoid: models that improve forecasting accuracy are systematically discarded.

5 Prediction evaluation as a model selection criterion

5.1 Some experiments with fixed coefficients

Two maps are generated from a simulation design that is similar to the one of Section 3. Data are generated from ARMA models with 100+100+10 observations. 1000 replications are conducted. Contrary to Figures 1–3, relative evaluation will be based on single trajectories. In MSE evaluation, prediction errors from 10 moving forecasts are squared and averaged. The model with the lowest MSE is then selected as the most appropriate prediction tool. In AIC evaluation, a single AIC is calculated from observations #101–#200 for the three models, i.e., the autoregressive, the moving-average, and the ARMA model. The model with the lowest AIC is then selected as the optimum time-series description.

Figure 7 shows the relative selection frequency of the pure autoregressive model on the basis of the smallest RMSE as calculated from only ten out-of-sample single-step forecasts. For both θ and ϕ small, the frequency is around 0.2 and gradually increases as $\phi \uparrow 1$. For $\theta = 1$, selection frequency reaches a minimum, excepting $\theta = \phi = 1$. It is interesting to note that the maximum frequency of selecting the autoregressive structure is not obtained for $\theta = 0$ but on a skew 'ridge' that runs from $(\phi, \theta) = (0, 0.5)$ to $(\phi, \theta) = (1, 0)$. The autoregressive model also dominates for $\theta = \phi = 1$, probably due to the unsatisfactory performance of the maximum-likelihood estimator for the ARMA(1,1) model in

this region.

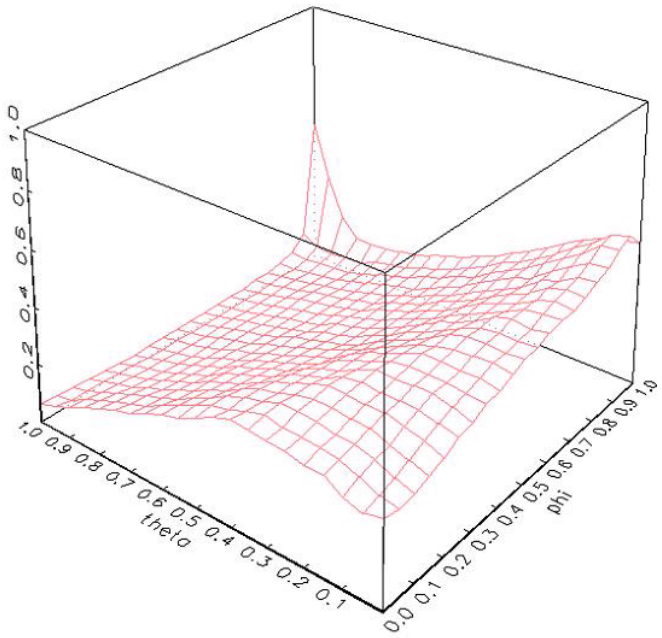

Figure 7: Prediction MSE as a model selection criterion: relative selection frequency of the AR(1) model when the data are generated by ARMA(1,1) models.

Figure 8 gives the relative selection frequency of the pure moving-average model. The moving-average model is unlikely to be selected for $\phi = 1$, where its frequency drops to zero, while it is selected in more than 40% of the cases with $\phi = 0$. As for the autoregressive model, selection frequency is not entirely monotonously dependent on ϕ, however.

Figure 9 gives the relative selection frequency of the mixed ARMA(1,1) model. The most sophisticated model has a low probability of being selected for $\phi = 0$ or for $\theta = 0$, albeit a slightly higher one for near white noise. This probability rises as ϕ and θ increase, excepting the area around $\phi = \theta = 1$, with the already mentioned problems of the estimation procedure.

The true information criteria permit a more precise model selection than the RMSE evaluation, as can be seen from Figures 10–12. There is a 100% preference for the mixed model when both ϕ and θ are large enough, and there is a strong preference for the autoregressive and moving-average structures close to their natural habitats. The poor performance of the ML estimator close to the point $\phi = \theta = 1$ is confirmed by the peak in that corner in Figure 10 and the corresponding trough in Figure 12.

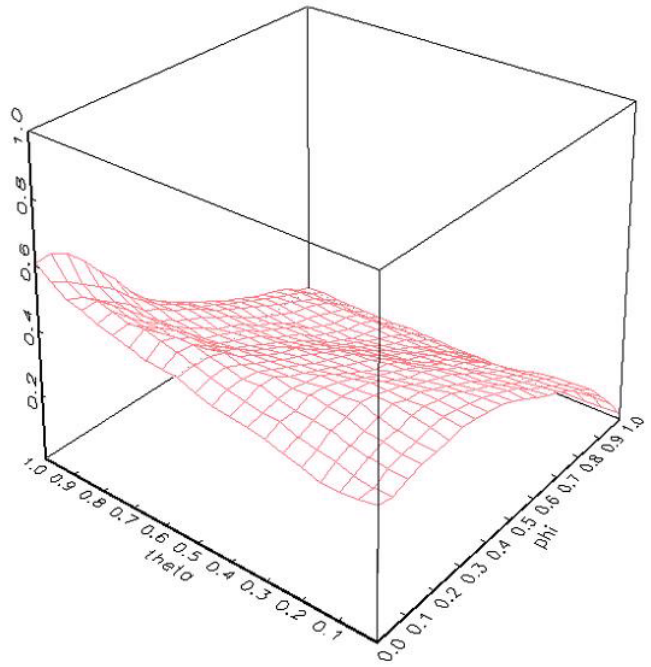

Figure 8: Prediction MSE as a model selection criterion: relative selection frequency of the MA(1) model when the data are generated by ARMA(1,1) models.

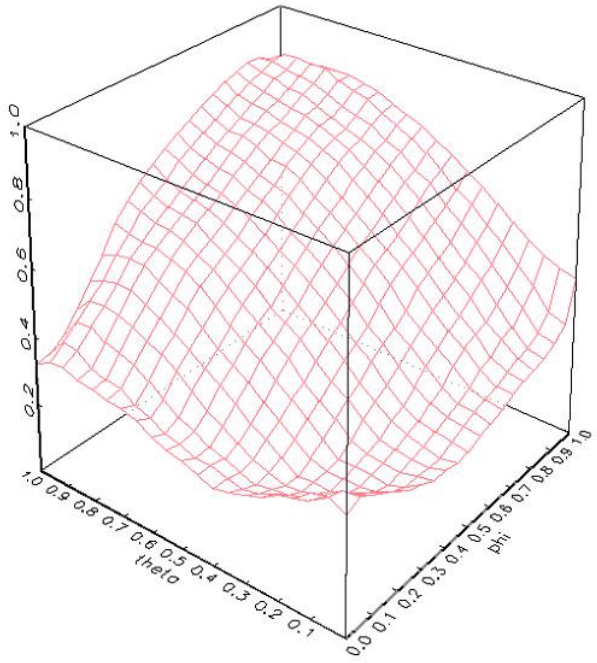

Figure 9: Prediction MSE as a model selection criterion: relative selection frequency of the ARMA(1,1) model when the data are generated by ARMA(1,1) models.

The experiment was repeated for the Schwarz criterion in place of the AIC. Figures of selection frequencies turned out very similar to Figures 10–12 and are therefore not shown. It appears that the recent statistical literature generally prefers AIC and its variants to BIC, which, despite its asymptotic consistency property, does not select satisfactorily in small samples. This paper focuses exclusively on the AIC.

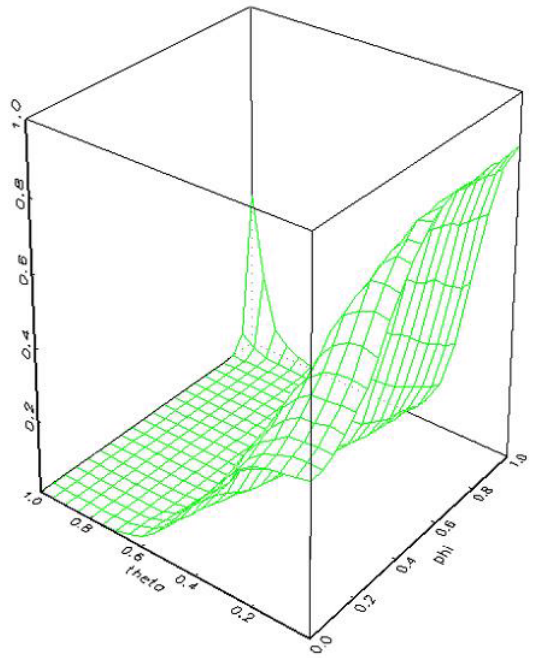

Figure 10: AIC as a model selection criterion: relative selection frequency of the AR(1) model when the data are generated by ARMA(1,1) models.

The conclusions from these experiments are twofold.

Firstly, comparative MSE or other forecast precision evaluations are poor substitutes for traditional information criteria when it comes to model selection. While the MSE graphs were based on trajectories of length 110, the AIC graphs used a length of 100 throughout. Also, computing time for the AIC graphs was approximately 5% of the computing time of the MSE graphs. Nonetheless, visual impression clearly supports model selection by AIC, provided one is interested in finding the 'true' model. If the aim is choosing a prediction model, the true model is not necessarily the better choice and the MSE–selected model, although incorrect, may dominate its AIC–selected counterpart. Also, one should keep in mind that, in practical applications, the data-generating process may be much more complex than the entertained model classes. The main relative weakness

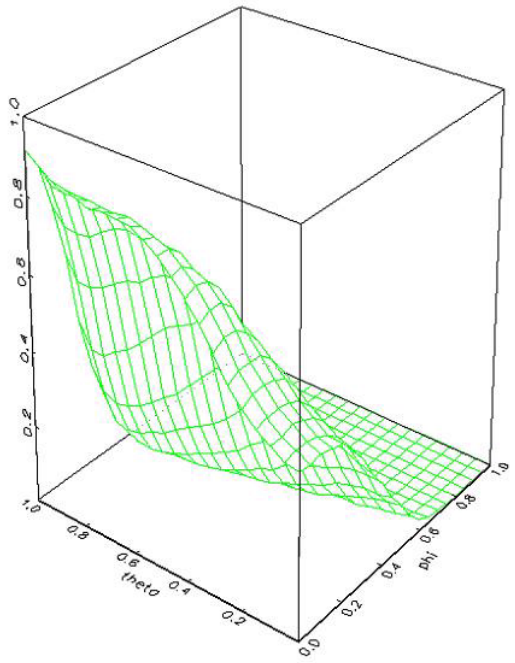

Figure 11: AIC as a model selection criterion: relative selection frequency of the MA(1) model when the data are generated by ARMA(1,1) models.

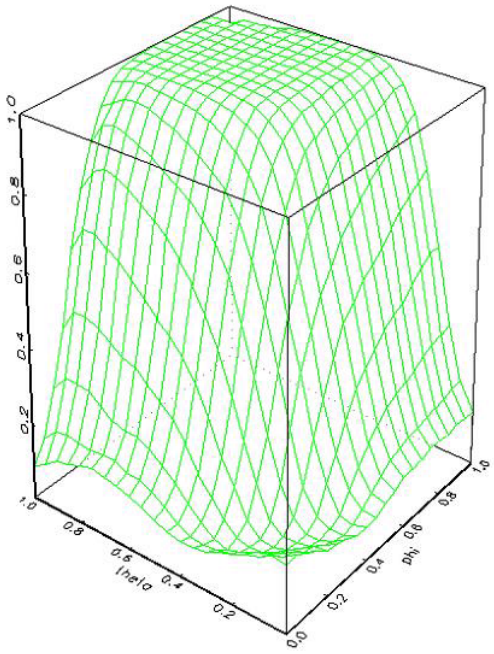

Figure 12: AIC as a model selection criterion: relative selection frequency of the ARMA(1,1) model when the data are generated by ARMA(1,1) models.

of the MSE criterion is its exclusive reliance on few observations at the end of the sample. This weakness may turn into a virtue if all entertained models are poor approximations at best and, for example, an ARMA(1,1) structure with time-changing coefficients approximates better than a linear one. This point is taken up in Section 5.2.

Secondly, however, the figures demonstrate that the recommendations by MSE comparisons are altogether similar and comparable to the AIC recommendations. In other words, an MSE comparison is nothing else than an evaluation of model selection statistics. If researchers are reluctant to study the significance of AIC decisions, there is no reason for a demand for significance tests on MSE or MAE decisions.

Finally, one should not overlook an important argument that may favor MSE/MAE comparisons over IC. While forecasting evaluations automatically penalize model complexity in such a way that the optimum prediction workhorse is supported, IC evaluations solely rely on the number of parameters as a sometimes poor measure of complexity. The traditional IC approach may work for choosing among simple linear structures but it may fail for nonlinear models with high inherent complexity and comparatively few parameters. Forecasting comparisons will continue to give the appropriate implicit penalty to the inconvenient nonlinear workhorse.

An important argument is certainly that model selection may not be the ultimate aim, particularly when forecasting criteria are considered. The shown graphs only demonstrate that AIC outperforms MSE with regard to the search for the 'true' structure. They do not show the relative prediction performance of the selected models. Therefore, another set of simulations were conducted, using the same ARMA models to generate trajectories as before, although now of length 211 instead of 210. Three model selection procedures are considered:

1. Models are selected on the basis of AIC. Here, 1000 replications are generated. For each replication, the selection based on observations #100–#200, on #101–#201 etc. is evaluated. The model with minimum AIC is used as a prediction model for the next observation, that is, for #201, #202, ... For each of the 10,000 cases, the MSE and MAE are collected. Finally, MSE (and MAE) are averaged over 10,000 cases.

2. Models are selected on the basis of MSE. For each replication, first the observation #200 is forecasted using #100–#199. The model with minimum MSE is then used to predict observation #201 from #100–#200. Then, the end point of the sample is shifted, until ten forecasts are available. The implied MSE is averaged over 10,000 cases.

3. First, each of the three models generates forecasts over a training period #201–#210. From these ten forecasts, the DM statistic is calculated to compare the basic AR(1) prediction and the more 'sophisticated' MA(1) and ARMA(1,1) rivals. If the rivals achieve a 'significantly' better performance at the one-sided 2.5% risk level, they are used in order to predict

#211. Otherwise, prediction relies on the AR(1) model. The resulting MSE is averaged over 1000 cases.

While the resulting model selection frequencies for the two former cases have already been reported, those for the latter case are shown in Figures 13–15. The default model of AR(1) is selected with a large probability over the whole parameter space, with a minimum for generating models with a large MA component. The default model is rarely overruled by the pure MA(1) model, even when θ is large. Mixed ARMA(1,1) are increasing in their selection frequency as ϕ and θ increase, although they never become as dominant as for the other model selection techniques.

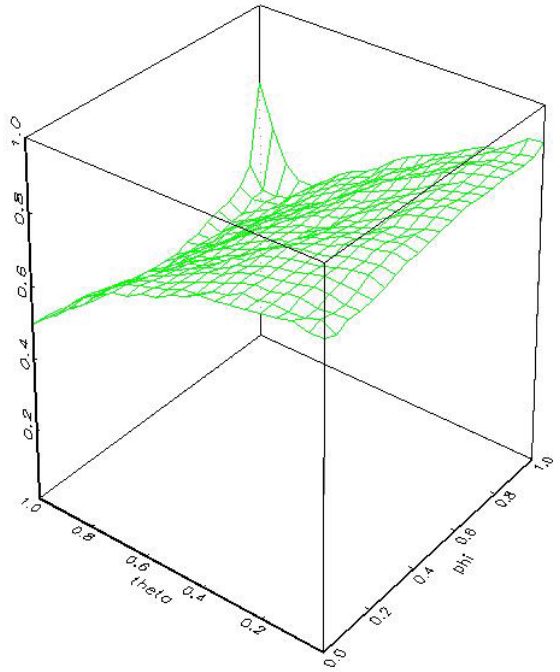

Figure 13: MSE plus DM significance as a model selection criterion: relative selection frequency of the AR(1) model when the data are generated by ARMA(1,1) models.

Again, it appears that model selection guided by MSE and an additional DM step does a poor job. Still, it may do a better job if forecasting is the ultimate aim of the exercise. Figure 16 shows that strategy #2, i.e., model selection based on the MSE, indeed achieves better forecasts than strategy #1 in a portion of the parameter space. The portion is characterized by small ϕ and large θ values, where AIC correctly chooses ARMA models but the bias

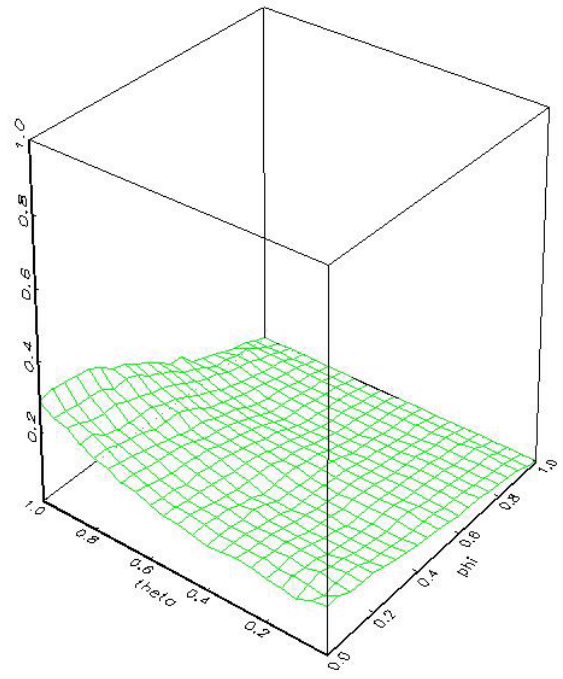

Figure 14: MSE plus DM significance as a model selection criterion: relative selection frequency of the MA(1) model when the data are generated by ARMA(1,1) models.

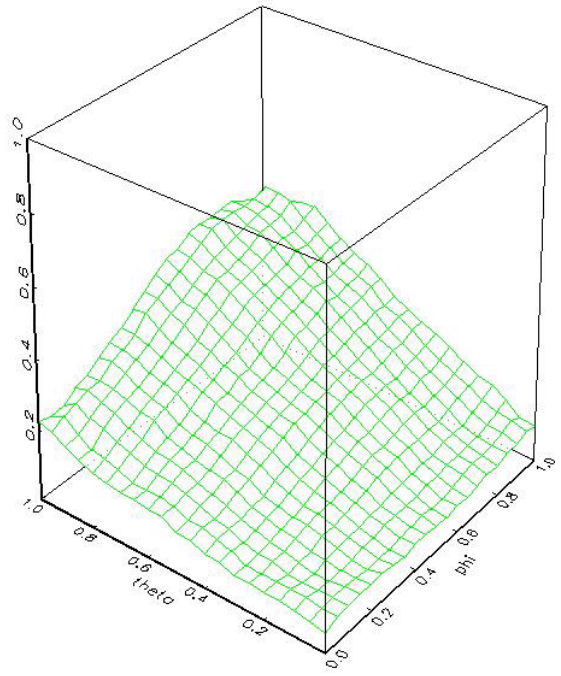

Figure 15: MSE plus DM significance as a model selection criterion: relative selection frequency of ARMA(1,1) when the data are generated by ARMA(1,1) models.

in ϕ estimation tends to outweigh the benefits from estimating it instead of restricting it at zero. For other (ϕ, θ) constellations, strategy #2 is worse and it becomes definitely inferior as $\phi \to 1$.

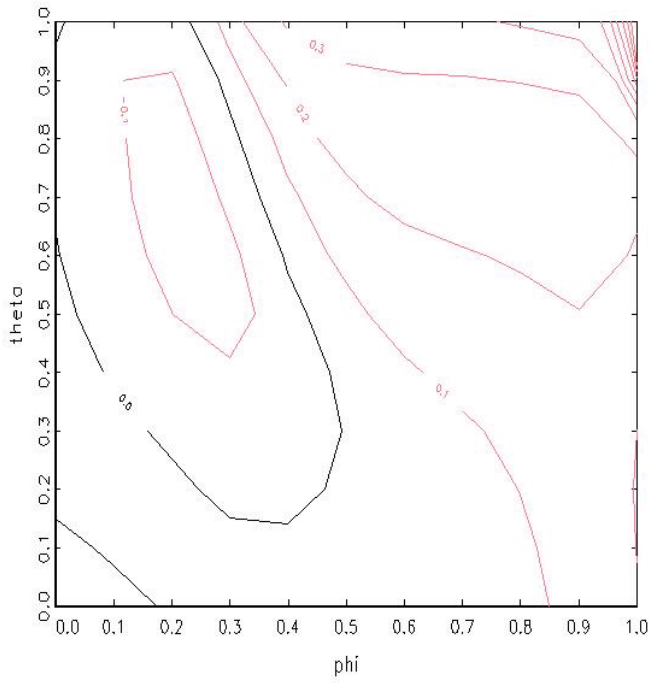

Figure 16: Contour plot of the difference of MSE achieved by MSE model selection minus MSE achieved by AIC model selection. 756 replications of $n = 100$ trajectories from ARMA processes.

Using the DM test on top of MSE selection generally succeeds in improving predictive accuracy relative to pure MSE selection. The reason is that the AR(1) forecasts are the most 'robust' predictions, while ARMA(1,1) forecasts bring in more bias and sampling variation. Figure 17 shows, however, that the MSE-DM search strategy is unable to beat AIC selection for those parameter values where pure MSE search has the most severe deficiencies, that is, for large ϕ. Standard AIC model selection achieves the best forecasting performance for $\phi > 0.5$, ignoring the north-east corner $(\phi, \theta) = (1, 1)$, a severely unstable model where the ARMA(1,1) estimator faces problems and it can be improved upon by replacing it by the AR(1) estimator.

We note that strategy #1 can be improved by replacing the asymptotic AIC criterion by the AIC_c suggested by Hurvich and Tsai, which indeed shrinks the area of MSE dominance considerably. Also, one may consider improving on strategy #2, replacing the selection based on a single forecast by the average

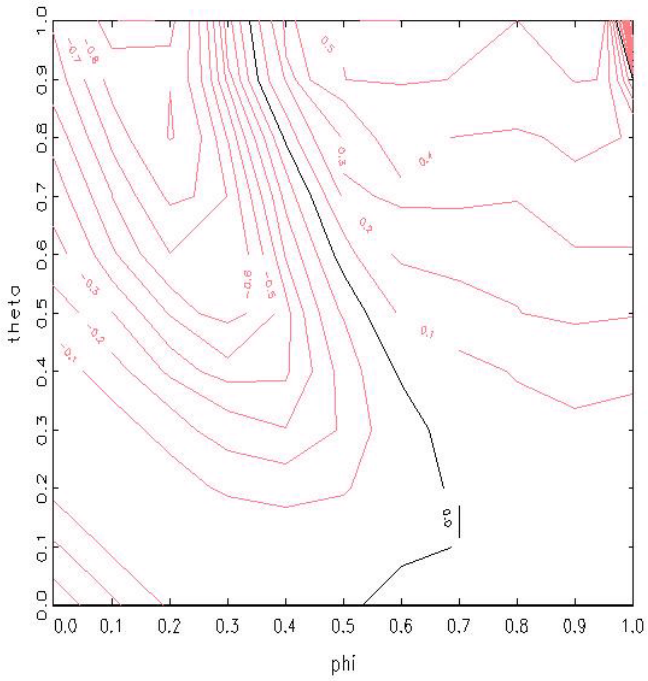

Figure 17: Contour plot of the difference of MSE achieved by MSE-DM model selection minus MSE achieved by AIC model selection. 756 replications of $n = 100$ trajectories from ARMA processes.

over a training period, as in strategy #3. In order to separate the effects of the DM test proper and of using a training set of 10 observations, the experiment was re-run with this modified strategy #2'. It turned out that the performance of strategy #2' does not differ substantially from strategy #2 and does not even dominate it over the whole parameter space. The slight improvement from strategy #3 relative to #2 mainly reflects its implicit preference for using autoregressive models in forecasting.

The bottom line is that DM testing and thus artificially biasing the decision toward a simple structure may help in cases where AIC selection chooses a correct model with unfavorable biases in parameter estimation. In those cases where AIC selection is to be preferred to MSE selection, DM testing does not help. All these remarks are only valid if the aim is prediction. If the aim is a search for the true model structure, out-of-sample prediction measures are much less likely to find that true structure than standard information criteria, while even this poor performance deteriorates if DM testing is applied.

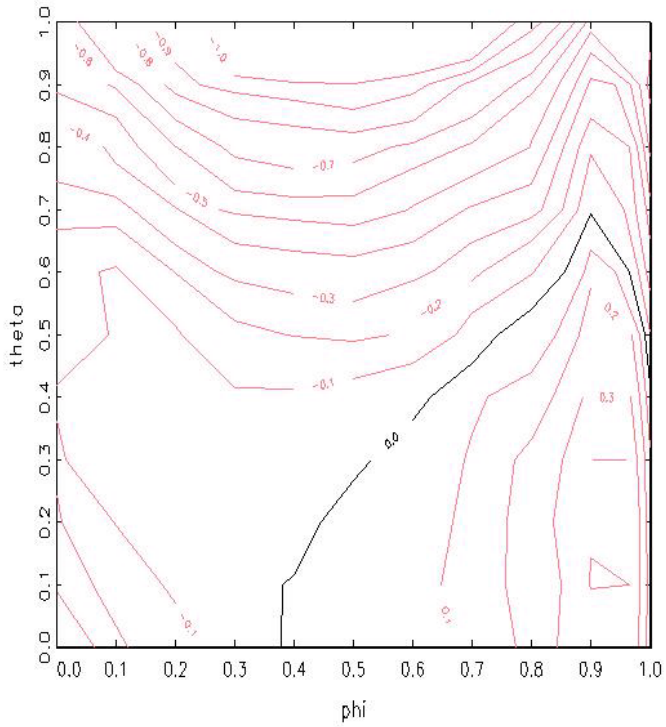

Figure 18: Contour plot of the difference of MSE achieved by MSE-DM model selection with MA default model minus MSE achieved by MSE model selection based on a 10–observations training set. 756 replications of $n = 100$ trajectories from ARMA processes.

A variant of this experiment, which is reported in less detail here, replaces the default AR model for DM testing by a default MA model. In this case, AIC yields better prediction for $\theta > 0.5$, while MSE–DM is preferable for $\theta < 0.4$. Figure 18 compares this moving-average DM strategy to the strategy #2'. DM testing incurs a deterioration for models with a dominant autoregressive component, i.e., for ϕ much larger than θ, while it incurs an improvement for other parameter values. The reason is similar to the one for the autoregressive version of the experiment. DM testing biases model selection more strongly toward choosing the MA structures, which is beneficial for MA–dominated generation mechanisms, although inappropriate for AR–dominated ones. When the true parameter value is unknown, the effects of the additional testing step are uncertain.

5.2 Some experiments with randomized coefficients

Data-generating models of the ARMA type may bias this research unduly toward model selection by IC. The sample size of 100 is large enough to allow reliable discrimination among AR, MA, and ARMA structures, accounting for the whole sample optimally. By contrast, model selection by forecasting criteria in the simulation design relies on the local behavior over a relatively short time span, whether the DM test is used or not. This observation motivates to investigate a different design that, conversely, may be biased in favor of forecasting criteria. Among economists, it is a common criticism of time-series models that the economic world is evolving and changing over time, hence any simple structure with constant coefficients may not be trustworthy. While it is not possible to generate models that match this idea of an evolving world in every aspect, ARMA processes with time-changing coefficients may be a good approximation. If the true structure is subject to changes with some persistence, forecasting based on a moving time window may capture the locally optimal prediction workhorse even when the assumed forecaster is unaware of the true data-generating process. What is otherwise an inefficiency, may be an asset of flexibility in such design. This design also meets the requirement that, in order to achieve a realistic design, the generating model should be more general than the structures in the set of entertained models.

Currently, arguments of coefficient variation often support the application of unobserved-components structures. However, most of these models are fixed-coefficients structures indeed, many of them assuming additional unit roots and very special short-run autocorrelation. By contrast, a model with truly changing structure is the ARMA model with stochastic coefficients, of the form

$$y_t = \phi_t y_{t-1} + \varepsilon_t + \theta_t \varepsilon_{t-1}. \tag{4}$$

In line with the time-series literature, such a model is called a RC–ARMA model (*random coefficients ARMA*). RC–ARMA models are simple generalizations of the RCA models studied by Nicholls and Quinn (1982). While the conditions for finite second moments are somehow stricter than in the constant-coefficients

ARMA model, strict stationarity even holds for $\mathrm{E}\,\phi_t > 1$, as long as the ϕ_t process is stable and its variation is not too large. Typical trajectories for $\mathrm{E}\,\phi_t$ equal or close to unity display short periods of very high or low values, before they turn back to longer periods of relatively small values. It should be noted that trajectories for $\mathrm{E}\,\phi_t = 0.9$, say, are quite different in appearance from ARMA(1,1) or AR(1) processes with $\phi = 0.9$. RCA models have been taken up in the literature on conditional heteroskedasticity in time series, and we refer to these publications for detailed conditions (see Tsay, 1987, Bougerol and Picard, 1992, Kunst, 1997).

In order to simplify this quite general structure somewhat, we convene that $\theta_t \equiv \theta$ but that

$$\phi_t - \phi = \zeta\left(\phi_{t-1} - \phi\right) + \eta_t, \tag{5}$$

with given $|\zeta| < 1$ and η_t *n.i.d.* $|\zeta| < 1$ is necessary to guarantee stability of the process (ϕ_t), as $\mathrm{E}\,(\phi_t - \phi)^2 = \left(1 - \zeta^2\right)^{-1} \mathrm{E}\,\eta_t^2$. For the reported simulation, $\zeta = 0.9$ is imposed, otherwise the standard design of the previous sections is used. $\mathrm{E}\,\eta_t^2$ is set at 0.1 times $\mathrm{E}\,\varepsilon_t^2$, which results in a sufficiently small variance for (ϕ_t) and in trajectories that are sufficiently different from constant-coefficients ARMA processes.

For this generating model, Figures 19–21 show the selection frequencies as surfaces if the MSE is chosen as the relevant model selection criterion. The graphs are not too different from the constant-coefficients case.

Using AIC as a model selector instead of the MSE, this time only based on the starting samples from observations #100–#200, results in the selection frequencies represented in Figures 22–24. Although the global maximum of selected AR models around $(\phi, \theta) = (0.6, 0)$ is to be noted, the general impression still favors the AIC selection. This impression changes, however, if the task of optimizing the forecasting properties is focused instead of selecting the true model structure, which is not exactly contained in the selection set anyway, as the generating models are not traditional ARMA models.

For the narrow-sense prediction evaluation experiment, the model selection decision was allowed to be revised as the sample end moved from #200 to #209. For each sample, a one-step prediction was calculated according to the AIC–selected model and according to the MSE–selected model. The difference MSE(MSE)-MSE(AIC) is shown as contour plots in Figure 25. For the MAE instead of the MSE, a very similar plot was obtained. The area $\{(\phi, \theta)\,|\phi > 0.7, \theta \in [0, 1]\}$ has been omitted, as these models show episodes of local instability that drive up measures such as MAE and MSE unduly. Generally, in these areas the differences are convincingly positive, which indicates that the AIC–selected model yields better predictions than the one determined by MSE optimization. Even within the shown area, dominance of prediction-gauged MSE selection is not as widespread as could have been expected. MSE selection performs better for structures that are close to white noise. Figure 21 shows that such structures are classified as ARMA(1,1) with positive probability, while AIC classifies them as simple AR(1) or MA(1). RC-ARMA models with small ϕ and θ are predicted better by ARMA(1,1) forecasts than by

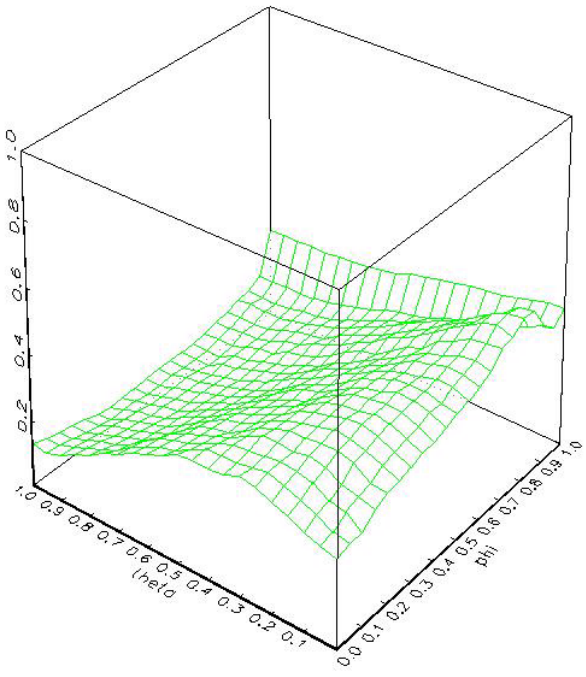

Figure 19: Prediction MSE as a model selection criterion: relative selection frequency of the AR(1) model when the data are generated by ARMA(1,1) models with random coefficients.

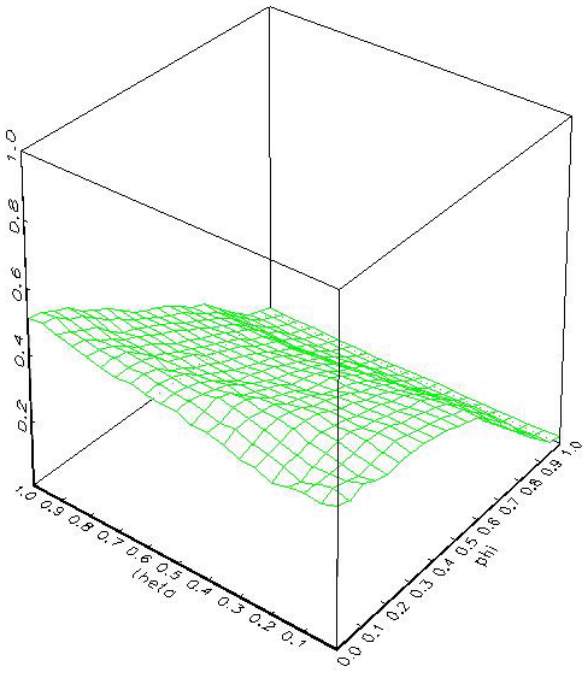

Figure 20: Prediction MSE as a model selection criterion: relative selection frequency of the MA(1) model when the data are generated by ARMA(1,1) models with random coefficients.

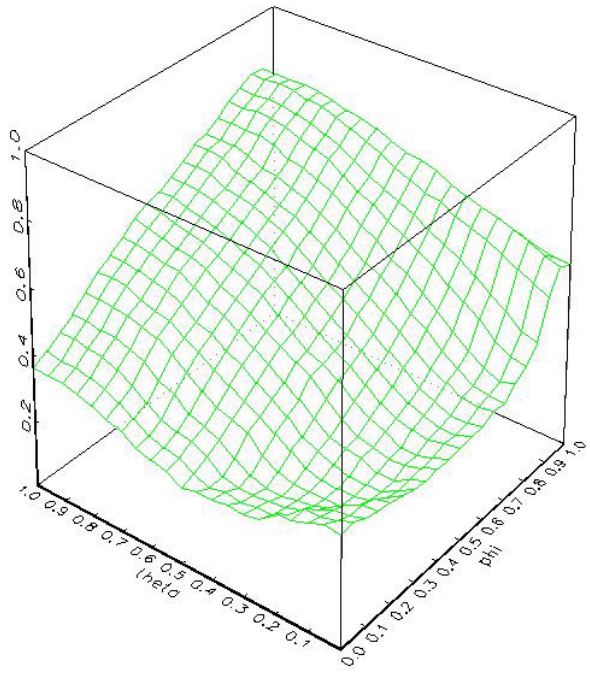

Figure 21: Prediction MSE as a model selection criterion: relative selection frequency of ARMA(1,1) model when the data are generated by ARMA(1,1) models with random coefficients.

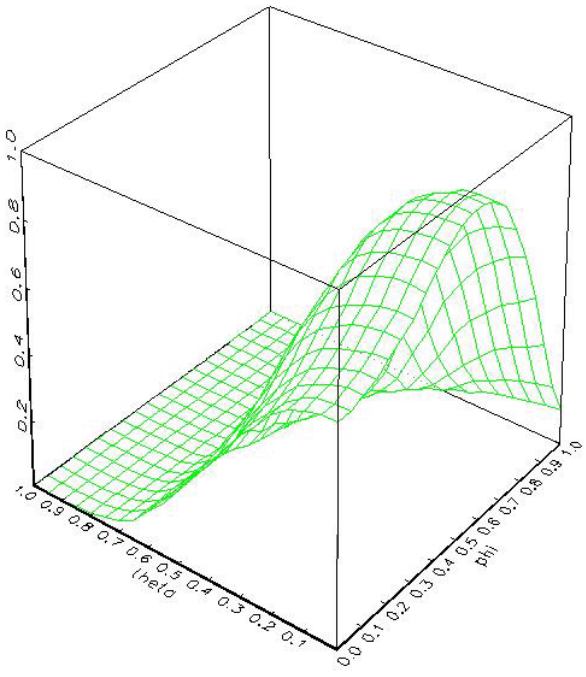

Figure 22: AIC as a model selection criterion: relative selection frequency of AR(1) model when the data are generated by ARMA(1,1) models with random coefficients.

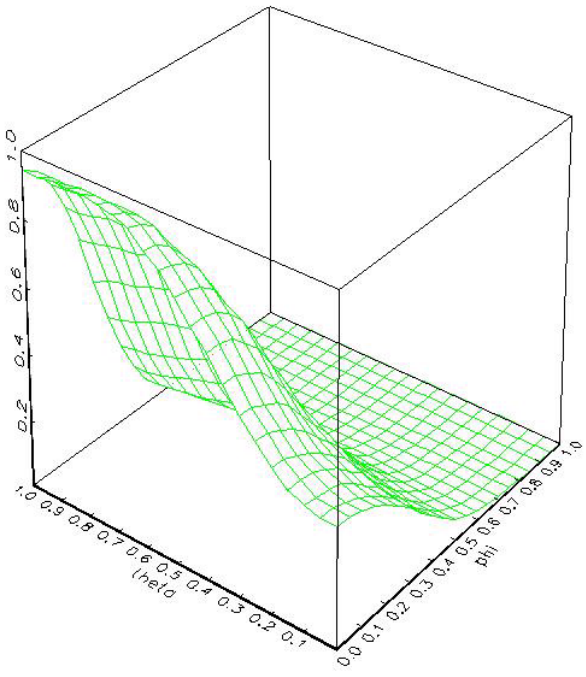

Figure 23: AIC as a model selection criterion: relative selection frequency of MA(1) model when the data are generated by ARMA(1,1) models with random coefficients.

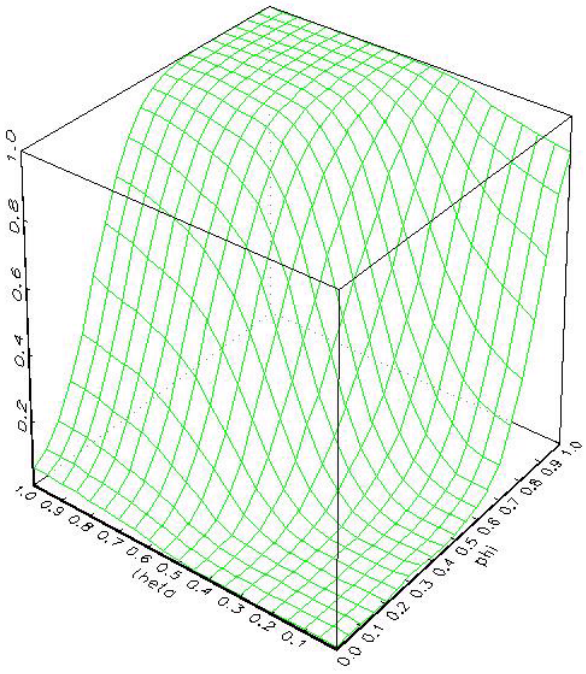

Figure 24: AIC as a model selection criterion: relative selection frequency of ARMA(1,1) model when the data are generated by ARMA(1,1) models with random coefficients.

constant-coefficients AR(1) or MA(1) forecasts, which can also be corroborated from other unreported simulation exercises. MSE selection also performs better for RC-ARMA models with small ϕ and large θ. AIC sees such trajectories as MA, while the MSE search picks up the evidence on a non-zero ϕ. For these structures, the criterion of parsimony may be a disadvantage. For larger ϕ, the MSE–selected models fall behind. The probability of achieving a smaller forecast by an AR or MA model is substantial, even when the trajectory at hand points to a full mixed ARMA. MSE search relies on this local evidence and inefficiently prefers the parsimonious structure, thus missing the chance to improve upon the prediction.

As a consequence of the experiments shown in Figure 25, 'selecting a forecasting model by forecasting' can be recommended for cases with low time correlation or comparatively short memory. Then, forecasting criteria may be able to exploit the advantage of adapting flexibly to time-changing coefficient structures. For larger time correlation, AIC dominates, as it gives a larger weight to the long-run information.

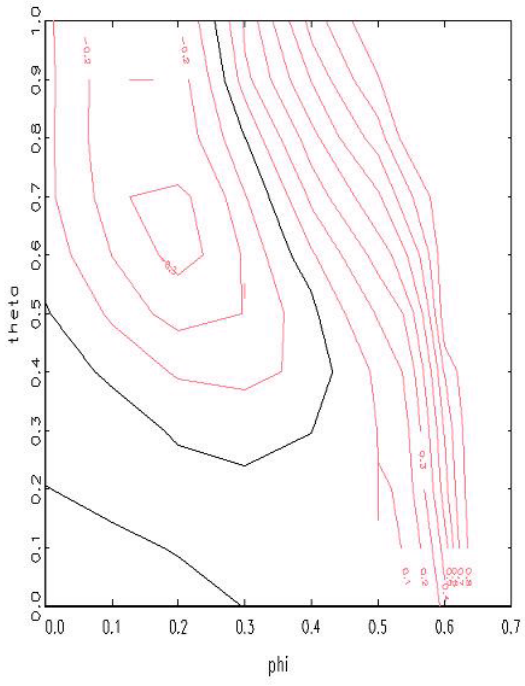

Figure 25: Difference of mean-squared prediction error of MSE–based model selection minus AIC–based model selection. Generating model is RC–ARMA(ϕ, θ).

Here, let us return to the main topic of this research. Suppose that the

forecaster optimizes the MSE but prefers the simplest model, as long as the rival models do not achieve a 'significant' improvement. Again, it appears natural to assume that the simplest forecasting model is the AR(1) model, which is typically preferred by most empirical researchers to the MA(1) and ARMA(1) models, as it can be estimated by least-squares regression and conveniently fits into the framework of distributed lags.

This experiment is comparable to the one reported above for the fixed-coefficients ARMA design. Applying the DM test to the very same experiment as before turned out to impose an excessive computational burden. Therefore, single-step predictions were evaluated for the range #201-#210, from which the DM statistic was calculated. Then, the AR(1) model acted as the default model, from which a forecast for #211 was used, unless either the MA(1) or the ARMA(1,1) model yielded a 'significantly' better prediction for the training range. If both models came out significantly better than AR(1), the model with the larger DM statistic was selected. Figures 26–28 show the selection frequencies for the three competing models. By construction, AR models are selected with a sizeable frequency over the whole parameter region, while MA models rarely are 'significantly' better than the AR baseline. As expected, mixed ARMA models attain their maximum frequency for generating structures with large θ and ϕ, though they are not as dominant for these values as for the other model selection strategies.

The results of this model selection procedure, which is based upon MSE and DM test significance, are reported as contour plots. Figure 29 gives the differences of the mean-absolute prediction errors for this complex procedure versus the AIC–based selection. AIC is preferable for larger ϕ, while MSE–DM is better for small ϕ and particularly for large θ. While this outcome is comparable to the relative performance of the selection algorithm without DM testing (Figure 25), the impression changes for mean-squared instead of mean-absolute errors, when the area of preference for the MSE–DM selection increases considerably. The reason is that MSE–DM selection avoids occasional large prediction failure for some extremely unstable trajectories with large ϕ, while AIC selection yields better prediction on average.

Of special interest is the comparison between the DM–based procedure and pure MSE optimization. In analogy to the results for the fixed-coefficients ARMA model, it turns out that the DM-testing step incurs an improved performance for the largest part of the parameter space. DM testing implies larger MSE, i.e., a deterioration in a region with small θ and ϕ around 0.5. For large ϕ, pure MSE optimization and MSE–DM yield very similar results and cause a slightly confusing contour plot, which is therefore not shown. The improvement achieved by the DM step is strongest for large θ and small ϕ, where AIC identifies MA models, while the MSE prefers AR or ARMA structures and DM testing gives an additional boost to the AR model with its most 'robust' forecasting performance.

Again, an intermediate experiment was conducted, with model selection determined by the average MSE over a training period of 10 observations, in order to find out how much of the change from MSE to MSE–DM is to be attributed

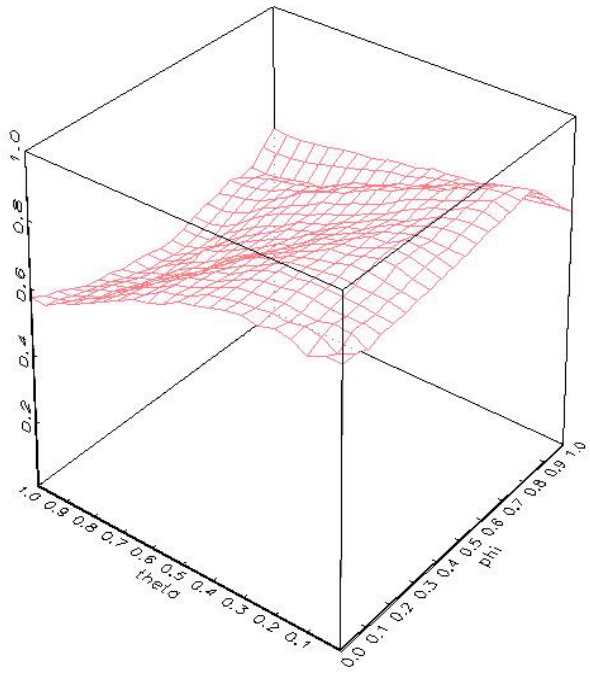

Figure 26: Selection frequency for AR(1) model according to a prediction evaluation using the MSE and the DM test over a training set of 10 observations, if the true structure is a RC–ARMA(1,1) model.

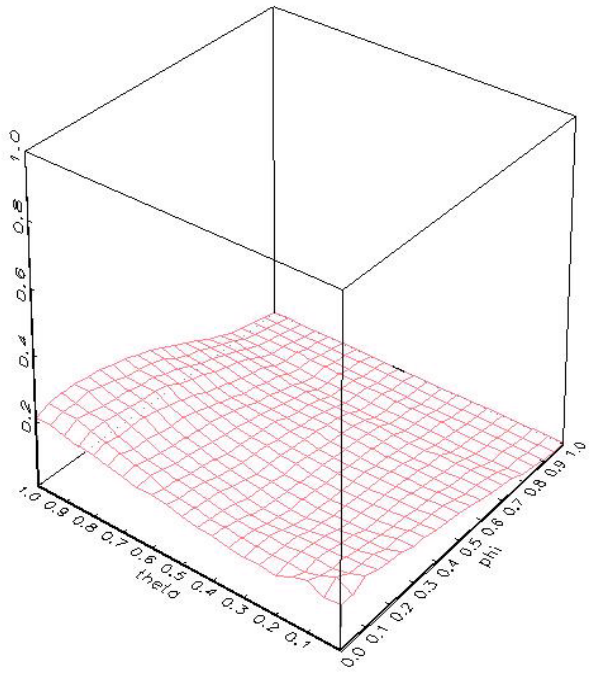

Figure 27: Selection frequency for MA(1) model according to a prediction evaluation using the MSE and the DM test over a training set of 10 observations, if the true structure is a RC–ARMA(1,1) model.

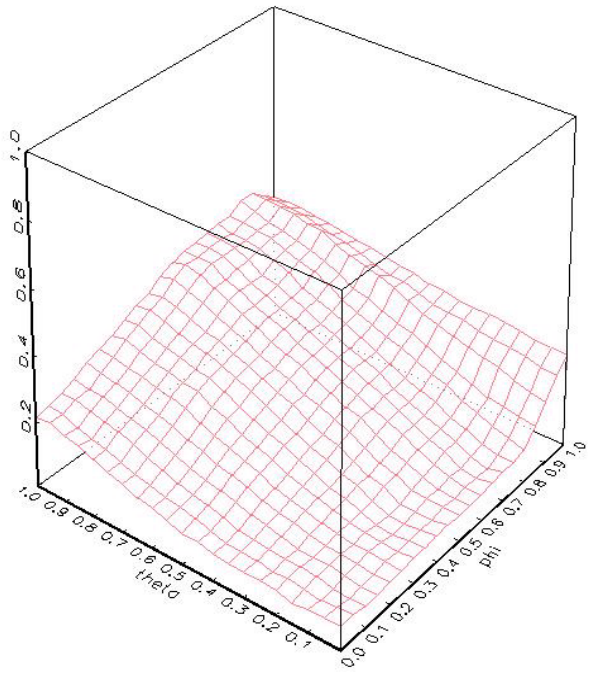

Figure 28: Selection frequency for ARMA(1,1) model according to a prediction evaluation using the MSE and the DM test over a training set of 10 observations, if the true structure is a RC–ARMA(1,1) model.

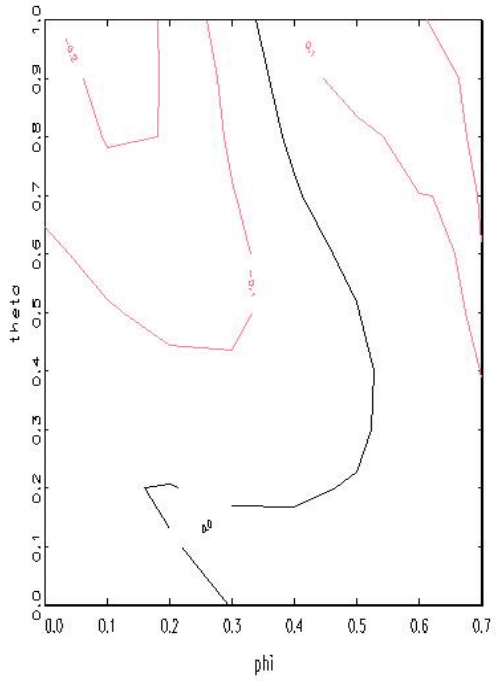

Figure 29: Difference of mean-absolute prediction error of MSE-DM–based model selection minus AIC–based model selection. Generating model is RC–ARMA(ϕ, θ).

to DM testing and how much to using a training period. For large ϕ, the intermediate design improves upon the single-observation MSE selection. For $\phi \geq 0.7$, averaging the MSE over a training phase even definitely beats MSE–DM. However, the change relative to single-observation MSE selection remains small, therefore the largest part of the displayed features are due to the DM testing step and not to the usage of a training phase. In order to focus on the main points, the results of the intermediate experiment are not shown.

The general conclusion to be drawn from the experiment is that application of the DM significance test offers little improvement for RC–ARMA data. The main drawback of MSE–based selection is that mixed ARMA structures are not recognized for large ϕ, due to the masking effects of the time variation in the coefficients and other models are selected instead, which results in bad prediction performance. This effect is even enhanced by the DM significance test that prevents discarding the AR model due to the relative merit of the full ARMA model not attaining 'significance' in the training period.

6 Summary and conclusion

Tests for the significance of differences in predictive accuracy have become a popular tool in empirical studies of predictive evaluation. Here, it is argued that the merit of such significance tests is doubtful. Theoretical arguments and simulation evidence is provided.

From a theoretical point of view, such tests are equivalent to testing the significance of a model selection decision. Like testing for significance on top of a model selection decision by AIC or by BIC, application of such a significance test is not justified and is usually motivated by a misunderstanding of model selection and a futile attempt of viewing information theoretic procedures in the framework of classical hypothesis testing. Such double checks implicitly give additional prior weight to the formal null hypothesis and, in practice, cause a bias in favor of simplicity, even though a penalty for complexity has been included already at the model selection stage.

From an empirical point of view, two simple, though empirically not implausible, model selection designs are evaluated by simulation. In the ARMA(1,1) design with fixed coefficients, choosing the AR(1) model as the 'null model' indeed points to a possible improvement of forecasting performance due to the secondary significance check. Searching for the optimum model by minimizing MSE over a relatively short training period indeed outperforms AIC search for sizeable parts of the parameter space. Both features are somehow surprising and may insinuate some more research by analytical and simulation methods. A possible reason for the second effect may be that AIC gives an insufficient penalty to the full ARMA(1,1) class in small samples. The effect can be mitigated by replacing the AIC by BIC or by the small-sample AIC_c criterion, as suggested by Hurvich and Tsai (1989). Another reason may be that, despite of the known advantageous forecasting properties of AIC–selected models, information criteria are still insufficiently tuned to the prediction task. With regard

to finding the true structure, AIC clearly dominates the rival procedures.

The other simulation example is a RC–ARMA(1,1) generating model with randomized coefficients. This more realistic design assumes that all models in the selection set are 'misspecified' and, instinctively, may favor MSE–based model search. However, the area of MSE dominance over AIC increases only slightly relative to the fixed-coefficients design. Subjecting this decision to a secondary significance test with an AR(1) null model causes an improvement in forecasting performance for random-coefficient MA models but a further deterioration for generating models with a large autoregressive coefficient.

It is obvious that it is fairly easy to find examples for data generation mechanisms that support any of the three considered selection procedures, hence this research is at a too early stage to allow specific recommendations. It is also obvious that each of the competing selection procedures can be modified and improved. Modified information criteria and modified versions of the original DM test have already been mentioned. However, the impression remains that significance testing for relative prediction accuracy is far from a panacea for the difficult task of improving forecasting accuracy in practice. The simulation evidence also points to the importance of separating the tasks of finding the optimum model as such in a Kullback-Leibler sense of approximating truth—and the task of optimizing predictive performance.

The author wishes to thank Jesus Crespo-Cuaresma for helpful comments. The usual provisio applies.

References

[1] Akaike, H. (1974) A New Look at the Statistical Model Identification . *IEEE Transactions on Automatic Control* AC–**19**, 716–723.

[2] Akaike, H. (1981) 'Modern development of statistical methods' Pages 169–184 in P. Eykhoff (ed.) *Trends and progress in system identification.* Pergamon Press.

[3] Ashley, R. (in press) 'Statistically significant forecasting improvements: how much out-of-sample data is likely necessary?' *International Journal of Forecasting*

[4] Bougerol, P., and Picard, N. (1982) 'Strict stationarity of generalized autoregressive processes' *Annals of Probability* **20**, 1714–1730.

[5] Box, G.E.P. and Jenkins, G. (1976) *Time Series Analysis, Forecasting and Control.* Revised edition, Holden-Day.

[6] Brockwell, P.J., and Davis, R.A. (2002) *Introduction to Time Series and Forecasting.* 2nd edition, Springer-Verlag.

[7] Burnham, K.P., and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* 2nd edition, Springer-Verlag.

[8] Diebold, F.X., and Mariano, R.S. (1995) 'Comparing Predictive Accuracy' *Journal of Business and Economic Statistics* **13**, 253–263.

[9] Fildes, R., and Stekler, H. (2002) 'The state of macroeconomic forecasting' *Journal of Macroeconomics* **24**, 435–468.

[10] Greene, W.H. (1990) *Econometric Analysis*. Macmillan.

[11] Hamilton, J.D. (1994) *Time Series Analysis*. Princeton University Press.

[12] Hurvich, C.M., and Tsai, C.L. (1989) 'Regression and time series model selection in small samples' *Biometrika* **76**, 297–307.

[13] Kunst, R.M. (1997) 'Augmented ARCH models for financial time series: stability conditions and empirical evidence' *Applied Financial Economics* **7**, 575–586.

[14] Linhart, H. (1988) 'A test whether two AIC's differ significantly' *South African Statistical Journal* **22**, 153–161.

[15] Nicholls, D.F., and Quinn, B.G. (1982) *Random coefficient autoregressive models: an introduction*. Lecture Notes in Statistics Vol. 11, Springer.

[16] Ramanathan, R. (2002) *Introductory Econometrics with Applications*. 5th edition, South-Western.

[17] Schwarz, G. (1978) 'Estimating the dimension of a model' *Annals of Statistics* **6**, 461–464.

[18] Tashman, L.J. (2000) 'Out-of-sample tests of forecasting accuracy: an analysis and review' *International Journal of Forecasting* **16**, 437–450.

[19] Tiao, G.C., and Tsay, R.S. (1984) 'Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models' *Journal of the American Statistical Association* **79**, 84–96.

[20] Tsay, R.S. (1987) 'Conditional heteroscedastic time series models' *Journal of the American Statistical Association* **82**, 590–604.

Author: Robert M. Kunst

Title: Testing for Relative Predictive Accuracy: A Critical Viewpoint
Reihe Ökonomie / Economics Series 130

Editor: Robert M. Kunst (Econometrics)
Associate Editors: Walter Fisher (Macroeconomics), Klaus Ritzberger (Microeconomics)

ISSN: 1605-7996

Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 59991-555 • http://www.ihs.ac.at

ISSN: 1605-7996